Predicting fake job posting using Machine Learning

Chetan sharaf
MT20109
chetan20109@iiitd.ac.in

Vandana sharma
MT20106
vandan20106@iiitd.ac.in

Rishabh bafna
MT20118
rishabh20118@iiitd.ac.in

Abstract

As with increase in digitisation of the society, every thing is happening digitally.And because of that the number of cybercrimes have increased, specifically in the area of job postings. So in this paper we are trying to show how machine learning algorithms can be used to predict fake job postings. Here we will be applying our models on the dataset of emails.Now as due to very less amount of fradulent data, oversampling was applied using imblearn library. Text mining procedures like counter vectorization and TFIDF was used for data processing. Various machine learning models were implemented and compared like, Mulitnomial Naive Bayes, Logistic Regression, Support Vector classifiers and Neural network model and a Deep learing model consisting of Long Short term Memory(LSTM).

1 Introduction

With ever increase in the data flow, everything in the world is becoming digital. Because of this there is sever rise in the online job market , where from job posting to working in the company is done online. And followed by this there are many fradulent job postings as well.

So in this paper we focus on the job posting part, where an effort is made to predict wheter a job posting floated out is real or fake. Although many work has already been done on this, we have employed various machine learning models like Mulitnomial Naive Bayes, Logistic Regression, Support Vector classifiers and Neural network model with different embedding techniques like Count vectorisation and Term-Frequency and Inverse Document Frequency(TF-IDF), which will make it easier for the machine learning model to run and predict.A deep learing model i.e LSTM was applied and compared so that how the dataset performs on a simple Deep learning framework

The major problems faced in most of the work that is previously done is seen is that almost all of them did not have good amount of data or did not had clean data. Here too we have sever data deficiency for the fake job posting but have enough data for Real job posting, which was overcame by oversampling. But due to lack of fake job posting data, the accuracy of the models came down.

2 Litreature Review

We have seen that in (Dutta and Bandyopadhyay, 2020), authors have considered mainly two types of classifier (Single classifier and ensemble classifier), for fraudulent job posts given out.For Single classifier based prediction they have used various form of classifier like Naive bias, Muli-layer perceptron, K-Nearest neighbour and Decision tree, where Decision Tree gives out the highest accuracy. And in Ensemble classifier , we have used Random forest, adabosster and gradient boosting methods , where gradient boosting methods gives out the highest accuracy of 97.65%. And Ensemble gives out the best value of accuracy and various other measurement standards.

In (Nasser and Alzaanin, 2020), various classifier modals were compared among Multinomial Naive bayes, SVM,KNN and decesion trees were compared , using TF-IDF text processing method and various standard evaluations methods were used and compared among them. Comaring all of them the paper suggested that Random forest will provide the highest accuracy, precision and F-measure while recall being one of the lowest.

In (Reis et al., 2019) , a special approach has been taken into consideration where the authors have hand crafted the features, based on the social media responses like biasness towards a perticular politaical party, as the data set taken is fake news from the social media.Various modals like KNN,

Naiev Bayes, random forest, SVM and XGB were taken and the results were analysed using AUC and F1 score. It was seen that Random forest and XGB perform the best among the following maodals taken. The handcrafted features might a bit difficult to implement as there might be interdepence between the features.

In (Androutsopoulos et al., 2000), two very basic methods were used for classifying wheter a perticular email is smap or not. The modals, Naive Bayesian Classifiacation and Memory based classification, with various other parameters changed, became pretty powerfull. The Naive Byaes performed comaritively well over 10 fold cross validation, and changing the number of attributes.

3 Data Analysis

Here we have taken data from kaggel(https://www.kaggle.com/shivamb/real-or-fake-fake-jobposting-prediction). The Dataset contains labeled datapoints and have 18 different features including the lable fradulent which is labeled to 0 for real job posting and 1 for fradulent job postings. The various features that we have are specified in Table 1. The Job id can be considered redundant as it provides no extra value to the dataset.

Features	Description
Job Id	Unique id for every job
Title	Title of the job
Location	Location of job
Department	Department of the job
Salary Range	Range of expected salary
Company Profile	Company description
Description	Details about the job
Requirements	work in company
Benefits	Benifits by the company
Telecommunicating	job in Telecomm?
Has Company logo	company has logo
Has Questions	1 if questions are present
Employment Type	Full time, part time etc
Required Experience	Previous experience
Required Education	Educational requirement
Industry	Type of industry
Function	Function of job
Fradulent	0 if not fraudulent else 1

Table 1: Features Table

In the dataset we have about 17880 different datapoints and about 866 of them are fradulent job posting, which might be a challenge as the number of fradulent jobs are much higher that the real jobs, so while training the model might simply be able to increase its accuracy by marking every test job as real. Also problem with the dataset is that the data of many features are missing is many instances, which can be seen from Figure 1

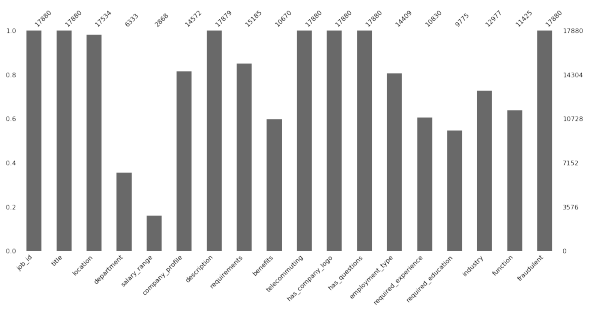

Figure 1: Number of available inastances in a feature

After analysing dataset , we can see that some of the feature are having much more focued on some of the instances, for example:

- Employment type - Full time (Figure 2)
- Required Experience - Mid-senior level,Entry level (Figure 3)
- Required Education - Bachlelor's Degree (Figure 4)
- Industry - Internet based job (Figure 5)

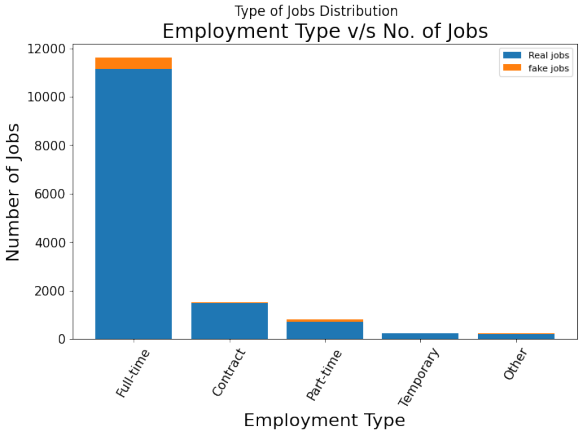

Figure 2: Employment type

3.1 Data-preprocessing

NLTK (Natural language Toolkit) is used in data processing. It is the main library for text data processing. Every Text base feature has been merge

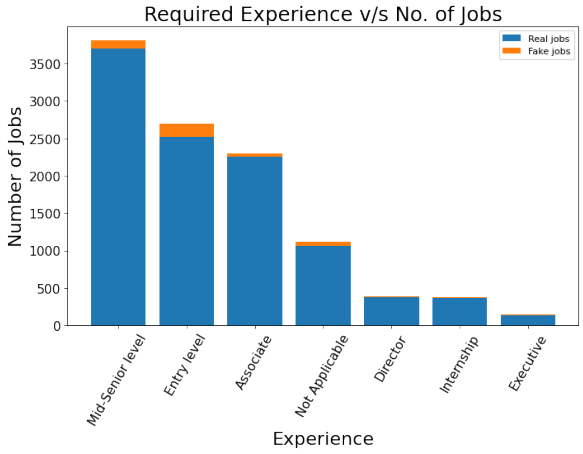

Figure 3: Required Experience

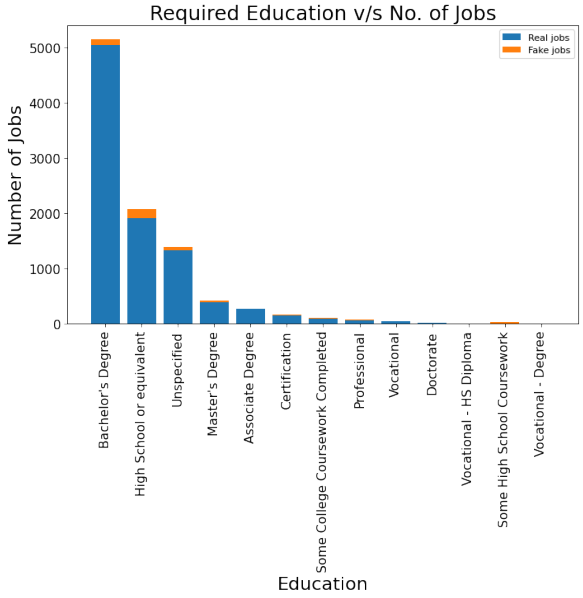

Figure 4: Required Education

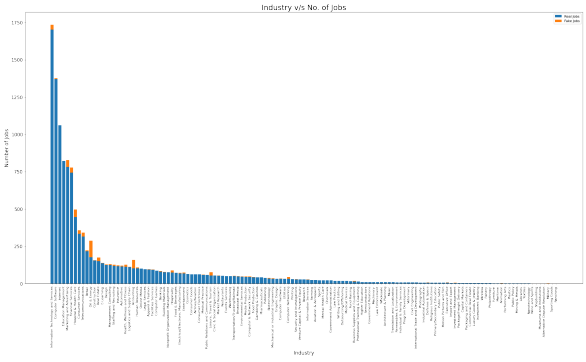

Figure 5: Industry

into one text feature so that counter vectorization and TFIDF can be done easily . Now as text based features has been merged so various unwanted features were deleted along with the job id as it bring nothing new to the table. The furthere processing was required ,so we took Tokenization into consideration , as it takes words separately into an array for analysis. The data was removed of various redundant words known as the stop words, These are the group of predefine words found from experimentation which does not give any meaningful analysis in the text. As the mail part also had various weblinks which might not bring any value for word embeddings so it was completly removed. Eg: in, and, the, which. Not required for analytics. For analyzing and the stopwords dictionary can be used and if any stop-word is not present in the Stopword dictionary then can pe added to the StopWord. Stemming and lemmatization are used for data preprocessing Stemming is defined as a stem is the base part of the word to which affixes can be attached for the derivatives. eg:"Combine" is the stem for combining, combine, and combined. Stemming is the process which converts a words into it's stem. Stemming keeps only the base word, thus reducing the total words in the corpus. Basically stemming cut off the affix of the word. Lemmatization it is Similar to stemming, but produces a proper root word that belong to the language. Eg: "Combine" is the lemmatized version of combine, combined and combining. Lemmatization uses a dictionary to match words to their root word. For data preprocessing the n-gram are used. n-grams is a sequence of n item in a sample text, Bigrams, trigrams, four-grams and so on.. Eg: "Dogs are favorite pets" Bigrams: (Dogs are), (are,favorite), (favorite,pets). Trigrams: (Dogs, are, favorite), (are, favorite,pets) As number of frauds are less then real jobs then that simply means that data is imbalanced so for balancing of data we have applied the Over Sampling. The need for oversampling can be seen by the Figure 6 After oversampling the follwing graph can be verified and is confirmed by the Figure 7

4 Results and discussion

For getting the results first we have splitted the data in 80:20 ratio for traing and testing. Two text mining procedures were applied which were, Counter vectorisation and TFIDF. Now as for applying data to a machine learing modal we need to convert text to numeric data which are handeld by these text mining procedures, as it becomes difficult to apply text based data to a machine learning modal.

For training the modal , various machine learn-

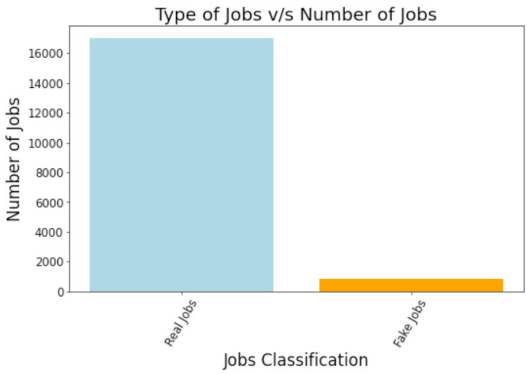

Figure 6: Job Classification

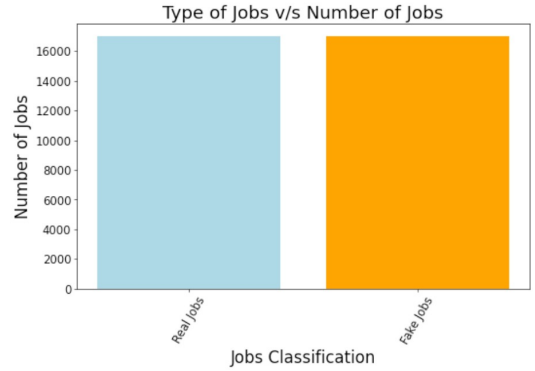

Figure 7: Post Sampling Job classification

ing models were used like Multinomial Naive Bayes, Logistic Regression , Support vector classifier and Multi-layer perceptron. All the Machine learning models were taken from Scikit learn library(an open source library), with the default setting. The resultant asccuracy can be seen from the Table 2

Models	CV	TF-IDF
MultinomialNB	96.22	96.00
Logistic Regression	98.04	97.95
Support Vector Classifier	98.07	98.60
MLP	98.60	98.76

Table 2: Result Table(Accuracy)

Just vewing the accuracy was not important , it was also important to note that the data was not balanced , so vewing individual F1 socre becomes significantly important to analyse the model , as sometimes even if the models mark Real for all the mails ,the accuracy can come out to be pretty high. So the F1 socres of all the machine learning models can be with real and fake jobs segregated can be seen from the Table 3.Where CV-R represents count vectorisation for Real values and CV-F represents count vectroisation for fake values. Analysing all the F1 Scores of the model it

Models	CV-R	CV-F	IDF-R	IDF-F
MNB	0.98	0.67	0.98	0.67
LR	0.99	0.78	0.99	0.78
SVC	0.99	0.78	0.99	0.81
MLP	0.99	0.83	0.99	0.85

Table 3: Result Table(F1 score for real and fake)

can be seen that Multilayed Perceptron i.e Neural Network performed the best with both Count Vectorisation and TF-IDF embedding. With TF-IDF the accuracy for Real values were 0.99 and for fake was 0.85. The Confusion matrix that we can see in the Figure 8 , clearly shows that the real was classified correctly almost everytime but there was significant error in classifying the fake jobs.

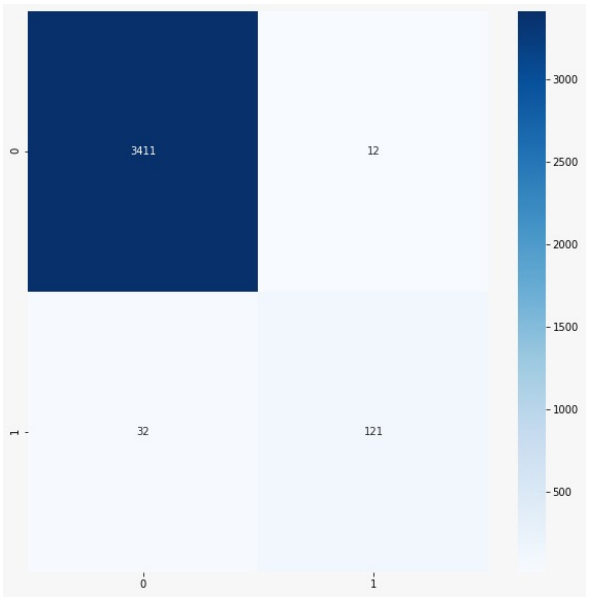

Figure 8: Confusion Matrix(MLP with TF-IDF)

A Deep learing model(LSTM) was also apllied with the embedding layer of the keras. And the following results were found. In this model a sngle layer of LSTM was used.

- Accuracy - 97.98
- F1 Score(Real) - 0.99
- F1 Score(Fake) - 0.74

The Confusion matrix of the model gives us a clear picture, which can be seen from the Figure 9

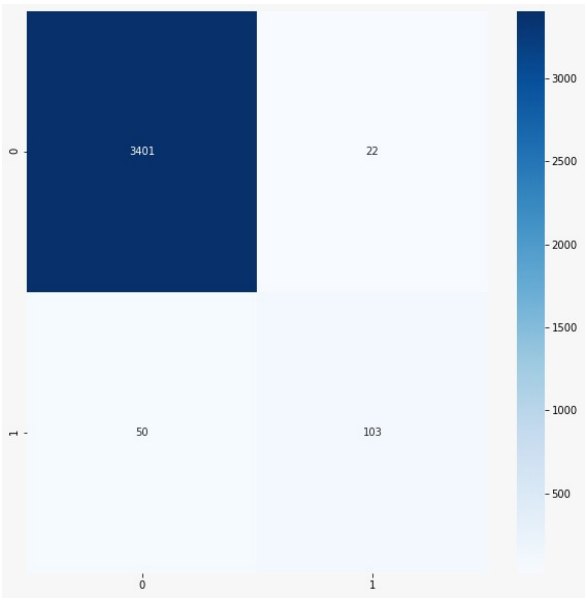

Figure 9: Confusion Matrix(LSTM)

5 Furthur work

The complete use of various supervised Machine learning algortihm was used on the mail data, but various features of the mail like location , department , salary range was removed from the usage for the simplification of the problem. So more features can be used to further improve the accuracy of the models.

References

Ion Androutsopoulos, Georgios Paliouras, Vangelis Karkaletsis, Georgios Sakkis, Constantine D. Spyropoulos, and Panagiotis Stamatopoulos. 2000. Learning to filter spam e-mail: A comparison of a naive bayesian and a memory-based approach. *CoRR*, cs.CL/0009009.

Shawni Dutta and Samir Kumar Bandyopadhyay. 2020. Fake job recruitment detection using machine learning approach. *International Journal of Engineering Trends and Technology*, 68(4):48–53.

Ibrahim M. Nasser and Amjad H. Alzaanin. 2020. Machine learning and job posting classification: A comparative study. *International Journal of Engineering and Information Systems (IJEAIS)*, 4(9):06–14.

J. C. S. Reis, A. Correia, F. Murai, A. Veloso, and F. Benevenuto. 2019. Supervised learning for fake news detection. *IEEE Intelligent Systems*, 34(2):76–81.

A Supplemental Material

The Jupyter-Notebook file,the data, and the processing data file can be downloaded from the link

https://drive.google.com/drive/folders/1jSgJg_60kz-eYDbcYmyWKrxjr39ZX6Ow?usp=sharing